Your Style Your Identity: Leveraging Writing and Photography Styles for Drug Trafficker Identification in Darknet Markets over Attributed Heterogeneous Information Network

Yiming Zhang, Yujie Fan
Department of CSEE
West Virginia University, WV, USA
{ymzhang,yf0004}@mix.wvu.edu

Wei Song, Shifu Hou
Department of CSEE
West Virginia University, WV, USA
{ws0016,shhou}@mix.wvu.edu

Yanfang Ye*, Xin Li
Department of CSEE
West Virginia University, WV, USA
{yanfang.ye,xin.li}@mail.wvu.edu

Liang Zhao
Department of IST
George Mason University, VA, USA
lzhao9@gmu.edu

Chuan Shi
School of Computer Science
BUPT, Beijing, China
shichuan@bupt.edu.cn

Jiabin Wang, Qi Xiong
Tecent Security Lab
Tecent, Guangdong, China
luciferwang@tencent.com

ABSTRACT

Due to its anonymity, there has been a dramatic growth of underground drug markets hosted in the darknet (e.g., Dream Market and Valhalla). To combat drug trafficking (a.k.a. illicit drug trading) in the cyberspace, there is an urgent need for automatic analysis of participants in darknet markets. However, one of the key challenges is that drug traffickers (i.e., vendors) may maintain multiple accounts across different markets or within the same market. To address this issue, in this paper, we propose and develop an intelligent system named *uStyle-uID* leveraging both writing and photography styles for drug trafficker identification at the first attempt. At the core of *uStyle-uID* is an attributed heterogeneous information network (AHIN) which elegantly integrates both writing and photography styles along with the text and photo contents, as well as other supporting attributes (i.e., trafficker and drug information) and various kinds of relations. Built on the constructed AHIN, to efficiently measure the relatedness over nodes (i.e., traffickers) in the constructed AHIN, we propose a new network embedding model *Vendor2Vec* to learn the low-dimensional representations for the nodes in AHIN, which leverages complementary attribute information attached in the nodes to guide the meta-path based random walk for path instances sampling. After that, we devise a learning model named *vIdentifier* to classify if a given pair of traffickers are the same individual. Comprehensive experiments on the data collections from four different darknet markets are conducted to validate the effectiveness of *uStyle-uID* which integrates our proposed method in drug trafficker identification by comparisons with alternative approaches.

CCS CONCEPTS

• **Security and privacy** → **Web application security**; • **Networks** → *Online social networks*; • **Computing methodologies** → **Machine learning algorithms**.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, CA, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313537

KEYWORDS

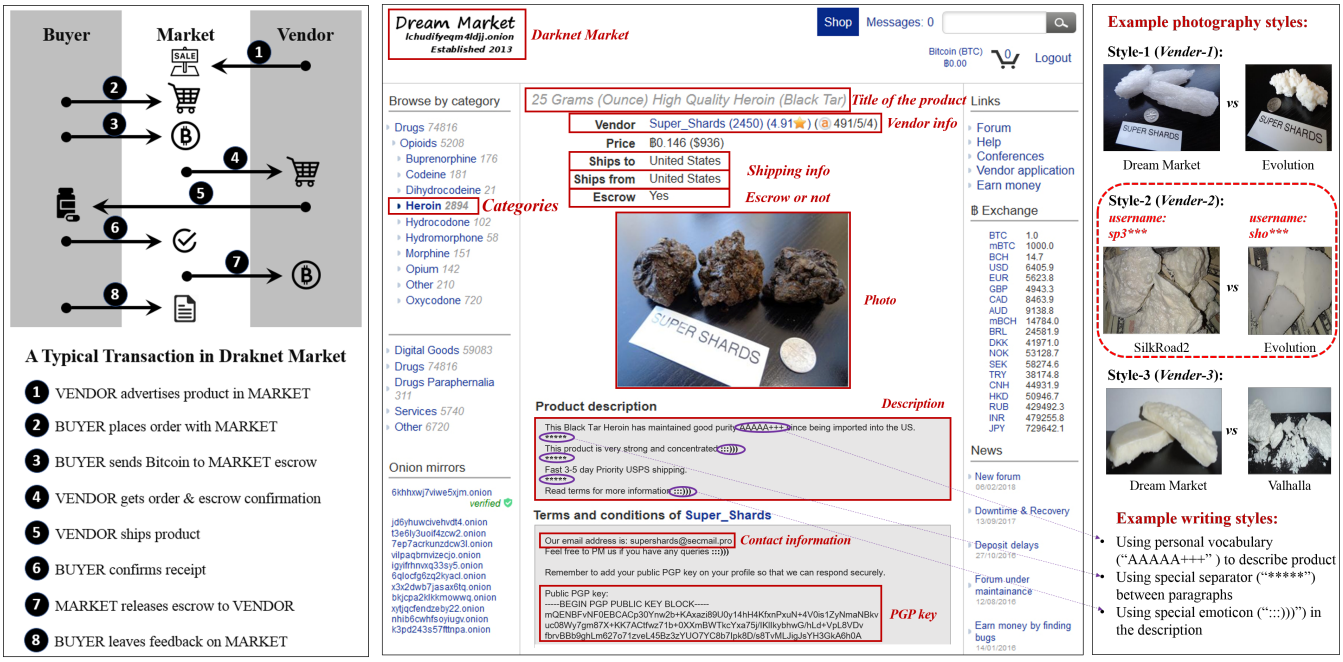
Darknet Market; Drug Trafficker Identification; Attributed Heterogeneous Information Network (AHIN); Network Embedding.

ACM Reference Format:
Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye*, Xin Li, Liang Zhao, Chuan Shi, and Jiabin Wang, Qi Xiong. 2019. *Your Style Your Identity*: Leveraging Writing and Photography Styles for Drug Trafficker Identification in Darknet Markets over Attributed Heterogeneous Information Network. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313537

1 INTRODUCTION

The market of illicit drugs (e.g., cannabis, cocaine, heroin) is considerably lucrative - i.e., the estimated yearly revenue for the global market reached about $426-$652 billion in 2017 [23]. Driven by such remarkable profits, the crime of drug trafficking (a.k.a. illicit drug trading) has never stopped but co-evolved along with the advance of modern technologies [4, 16, 17, 28, 32, 38, 39]. Darknet, as a hidden part of the Internet, employs advanced encryption techniques to protect the anonymity of its users. The markets hosted in the darknet are built on The Onion Router (TOR) service to hide the IP address, the escrow system, the encrypted communication tools like Pretty Good Privacy (PGP), and the virtually untraceable cryptocurrency (e.g., bitcoin) to facilitate anonymous transactions among participants. Figure 1.(a) illustrates a typical transaction in darknet markets. Due to its anonymity, there has been a dramatic growth of underground drug markets hosted in the darknet (e.g., Silk Road 3 [33], Dream Market [27], Valhalla [37], known as "eBay of drugs" or "Amazon of drugs"). Illegal trading of drugs in these markets has turned into a serious global concern because of its severe consequences on society (e.g., violent crimes) and public health at regional, national and international levels [8].

To combat drug trafficking in the cyberspace, there is an urgent need for analysis of participants in darknet markets, as it could provide valuable insight to the investigation of drug trafficking ecosystem and prediction of future incidents while building proactive defenses [6]. However, one of the key challenges is that drug traffickers may maintain multiple accounts across different markets or in the same market for the reasons [2, 4, 34] such as ripper (i.e.,



(a) A typical transaction in darknet market. (b) Illicit drug advertised by a trafficker (i.e., vender) in darknet market. (c) Illustration of writing & photography styles.

Figure 1: Illustration of drug trafficking in Darknet market.

an old account has lost the trust of other members), branding (i.e., a vendor creates an alias to positively review his/her own products or services), and anonymity. Linking different accounts to the same individuals is essential to track their status and better understand the online drug trafficking ecosystem [40]. Given the growing scale of darknet markets and the large number of user accounts, it is simply impossible to manually link suspicious accounts and track their latest status. Therefore, it is highly desirable to develop novel methodologies that can automatically link multiple accounts of the same individuals in darknet markets.

To automate the process, some of existing approaches [2, 5, 20] relied on stylometry analysis which aims at linking different accounts to the same user based on his/her writing styles (e.g., as illustrated in Figure 1.(b) and (c), the vendor likes using specific emoticon in the product description). Since drug traffickers in darkent markets have to prove the possession of illegal drugs by posting their own product photos, their distinct photography styles might be revealed by the posted photos. A recent research [40] proposed to link multiple accounts of the same vendors in different darknet markets based on their distinct photography styles (e.g., the way to display products and camera model as shown in Figure 1.(c)). Though each kind of analysis has shown success in fingerprinting underground market participants, using them respectively may suffer different challenges (e.g., stylometry analysis only is sensitive to the language of content [40], while photography style analysis only may face the challenge of intrinsic ambiguity arising from resale or photo plagiarizing). *Can we leverage both writing and photography styles to develop an integrated framework for drug trafficker identification in darknet markets?* To the best of our knowledge, this has yet studied in the open literature.

In this paper, we propose to leverage both writing and photography styles to develop an intelligent system (named *uStyle-uID*) to automatically link multiple accounts of the same individuals for drug trafficker identification in darknet markets. In *uStyle-uID*, given a pair of vendors (denoted by their usernames in the related markets), to determine whether they are the same individual, we not only analyze their posted contents (i.e., including their posted texts and photos), but also consider their writing styles and photography styles as well as other supporting attributes (i.e., vendor and drug information) and various kinds of relations. To depict vendors, drugs, texts, photos and their associated attributes as well as the rich relations among them, we present an attributed heterogeneous information network (AHIN) [26] for modeling. To tackle the challenge of high computation cost and memory constraint of measuring the relatedness over vendors in the constructed AHIN, we propose a new network embedding model named *Vendor2Vec* to learn low-dimensional attribute-aware embeddings for the nodes in AHIN. The proposed *Vendor2Vec* model leverages complementary attribute information of each node to guide the meta-path based random walk for path-instance sampling; then a skip-gram model [30] is utilized to learn effective node representations for AHIN. Finally, based on the learned latent representations of the nodes (i.e., vendors) in AHIN, we devise a learning model named *vIdentifier* to classify whether a given pair of vendors are the same individual.

2 PROPOSED METHOD

The overview of our developed system *uStyle-uID* for drug trafficker identification in darknet markets is shown in Figure 2. In this section, we will introduce the detailed approaches which are integrated in *uStyle-uID* for drug trafficker identification.

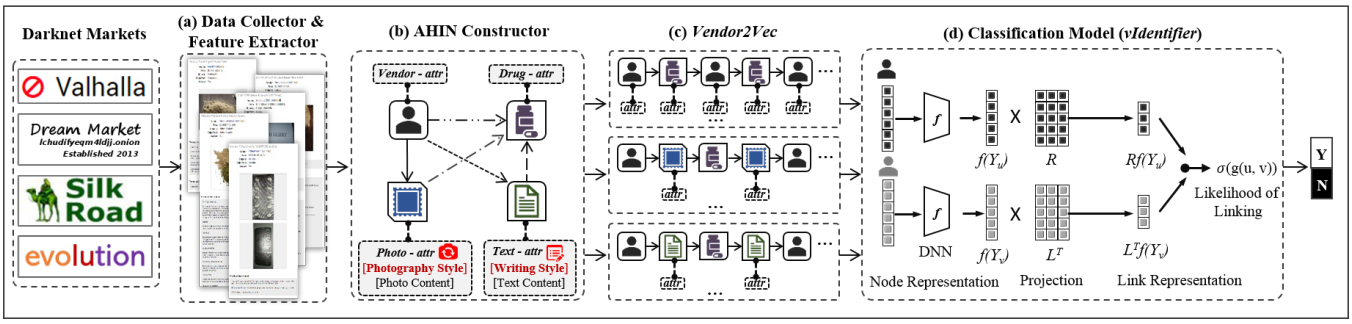


Figure 2: System architecture of *uStyle-uID*.

2.1 Feature Extraction

We propose to characterize vendors in darknet markets in a comprehensive view by extracting various features.

(1) Posted text and writing style extraction. To fingerprint a vendor based on his/her posted texts, we consider both his/her posted text content and writing style. For text content, we apply *doc2vec* [25] to convert each text of variant size into a fixed length feature vector (empirically, we set the dimension to 100). For writing style [2, 20], we propose to extract multi-scale stylometry features at three different levels as follows: 1) ***Lexical features*** can be further divided into character-based and word-based groups to capture stylistic traits. At this level, we extract *i*) number of characters, *ii*) number of digits/white spaces/special characters, *iii*) number of words, *iv*) average word length, and *v*) vocabulary richness [36]. 2) ***Syntactic features*** capture the writing style from the sentence structure. In this category, we adopt *i*) frequency of punctuation, *ii*) frequency of function word, *iii*) number of sentences beginning with a capital letter, and *iv*) frequency of parts-of-speech *n*-grams (we set $n = 3$ in our case). 3) ***Structural features*** represent the way an author organizes the layout of his/her posted text. We consider *i*) total number of paragraphs, *ii*) indentation of paragraph, *iii*) whether there's separator between paragraphs, and *iv*) number of words/sentences/characters per paragraph. For each posted text, we then concatenate its converted feature vector representing the posted text content and the feature vector describing its writing style as an attribute associated with this posted text.

(2) Posted photo and photography style extraction. To represent the content of a posted photo, we propose to utilize *image2vec* [15] to convert it into a fixed length feature vector (empirically, we set the dimension to 100). Since drug traffickers in darkent markets have to prove the possession of illegal drugs by posting their own product photos, their distinct photography style might be revealed by the posted photos. We propose to capture the photography style by extracting its *low-level* and *high-level* features. 1) ***Low-level features*** refer to the information that can be directly obtained from a photo's exchangeable image file format (EXIF) data, which include *i*) camera make and model, *ii*) camera angle, *iii*) exposure time, *iv*) focal length, and *v*) image size. 2) ***High-level features*** are extracted from the photo's original content. We first convert the photo into its HSV (hue, saturation, value) representation and then extract the following five types of high-level features: *i*) colorfulness, *ii*) exposure of light, *iii*) saturation, *iv*) hue count, and *v*) contrast. In our current implementation, colorfulness, exposure of light and saturation are calculated using the method in [7]; while hue count and contrast are measured by [24]. For each posted photo by a vendor, we then concatenate its converted feature vector representing the posted photo content and the feature vector describing its photography style as an attribute associated with this posted photo.

(3) Attributed features of vendors and drugs. Besides the above extracted features, vendors' basic information and drugs they sell also play an important role in resolving their identities. Therefore, we further extract three kinds of features to depict each vendor: *username*, *PGP key* and *contact information*. Note that, for username, we first apply standard string matching techniques to measure the similarity of two usernames, if their similarity is greater than a user-specific threshold, we regard these two usernames as the same (e.g., "MF***Jones" and "MF***J0nes"). For each drug, we further extract its *category*, *escrow information* and *shipping information* (e.g., from where and to where). Then, we apply one-hot encoding [41] to convert the extracted features to a binary feature vector to be an attribute associated with each vendor/drug.

(4) Relation-based Features. In order to characterize the rich relations among vendors, drugs, posted texts and photos, we further extract the following relation-based features: 1) ***R1***: the *vendor-sell-drug* relation indicates whether a vendor sells a drug; 2) ***R2***: the *vendor-write-text* relation denotes if a vendor writes a text; 3) ***R3***: the *vendor-post-photo* relation indicates whether a vendor posts a photo; 4) ***R4***: the *text-describe-drug* relation denotes whether a text describes a drug; 5) ***R5***: the *photo-characterize-drug* relation indicates if a photo characterizes a drug.

2.2 AHIN Construction

Though heterogeneous information network (HIN) [35] has shown the success of modeling different types of entities and relations, it has limited capability of modeling additional attributes attached to entities. Thus, to depict vendors, drugs, texts, photos and their associated attributes as well as the rich relationships among them, we propose to use attributed HIN (AHIN) for representation.

Definition 2.1. ***Attributed heterogeneous information network (AHIN)*** [26]. Let $\mathcal{T} = \{T_1, ..., T_m\}$ be a set of m entity types. For each entity type T_i, let $\mathcal{X}_i$ be the set of entities of type T_i and A_i be the set of attributes defined for entities of type T_i. An entity x_j of type T_i is associated with an attribute vector $\mathbf{f}_j = (f_{j1}, f_{j2}, ..., f_{j|A_i|})$. An AHIN is defined by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with an *entity* type

mapping $\phi: \mathcal{V} \rightarrow \mathcal{T}$ and a *relation* type mapping $\psi: \mathcal{E} \rightarrow \mathcal{R}$, where $\mathcal{V} = \bigcup_{i=1}^{m} \mathcal{X}_i$ denotes the entity set and $\mathcal{E}$ is the relation set, $\mathcal{T}$ denotes the entity type set and $\mathcal{R}$ is the relation type set, $\mathcal{A} = \bigcup_{i=1}^{m} A_i$, and the number of entity types $|\mathcal{T}| > 1$ or the number of relation types $|\mathcal{R}| > 1$. The ***network schema*** [26] for an AHIN $\mathcal{G}$, denoted by $\mathcal{T}_\mathcal{G} = (\mathcal{T}, \mathcal{R})$, is a graph with nodes as entity types from $\mathcal{T}$ and edges as relation types from $\mathcal{R}$.

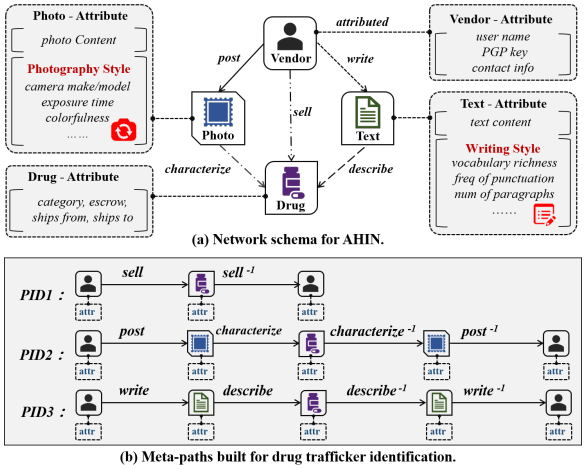


Figure 3: Network schema and meta-paths.

In our application, we have four entity types and five types of relations among them; meantime, each entity is also attached with an extracted feature vector representing its associated attributes. Based on the definitions above, the network schema for AHIN in our case is shown in Figure 3.(a) (to facilitate the illustration, attribute information is shown in its original form). Then, we adopt the concept of meta-path [35] to formulate higher-order relationships among entities in AHIN. In our application, we focus on three most meaningful meta-paths (i.e., ***PID1-PID3*** as shown in Figure 3.(b)) to jointly characterize the relatedness between two vendors from different views: (1) ***PID1*** means that two vendors can be connected through the path that they both sell the same kind of drug (e.g., heroin); (2) ***PID2*** denotes that two vendors can be linked if their posted photos describe the same kind of drug (e.g., as shown in Figure 1.(c), the vendor with username of "sp3***" in SilkRoad2 and the vendor with username of "sho***" in Evolution can be linked via this meta-path); (3) ***PID3*** indicates that two vendors can be connected if their written texts describe the same kind of drug.

2.3 *Vendor2Vec*

To reduce the computation and space cost in network mining, scalable representation learning method [10, 14] for AHIN is in need. We formalize the AHIN representation learning problem as below.

Definition 2.2. ***AHIN Representation Learning*** [10, 14]. Given an AHIN $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, the representation learning task is to learn a function $f: \mathcal{V} \rightarrow \mathbb{R}^D$ that maps each node $v \in \mathcal{V}$ to a vector in a D-dimensional space $\mathbb{R}^D$, $D \ll |\mathcal{V}|$ that are capable of preserving both structural and semantic relations among them.

To solve this problem, we propose a novel *attribute-aware* AHIN embedding model named *Vendor2Vec* which consists of attribute-aware meta-path random walk and skip-gram model. Given an AHIN $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with schema $\mathcal{T}_\mathcal{G} = (\mathcal{T}, \mathcal{R})$, and a meta-path scheme $\mathcal{P}$ in the basic form: $T_1 \rightarrow \cdots T_t \rightarrow T_{t+1} \cdots \rightarrow T_L \rightarrow T_{L+1}$, we use attribute-aware meta-path to guide a random walker in AHIN, the transition probability at step i is calculated as:

$$p(v^{i+1}|v_t^i, \mathcal{P}) = \begin{cases} \frac{sim(\mathbf{f}_{v'}, \mathbf{f}_{v^{i+1}})}{\sum_{v^c \in N_{T_{t+1}}(v_t^i)} sim(\mathbf{f}_{v'}, \mathbf{f}_{v^c})} \\ \qquad (v_t^i, v^{i+1}) \in \mathcal{E}, \phi(v^{i+1}) = \phi(v') = T_{t+1} \\ \frac{1}{|N_{T_{t+1}}(v_t^i)|} \\ \qquad (v_t^i, v^{i+1}) \in \mathcal{E}, \phi(v^{i+1}) = T_{t+1}, v' = \emptyset \\ 0 \qquad \text{otherwise,} \end{cases} \tag{1}$$

where v' denotes the latest entity the walker visited is with the same type of v^{i+1}, $sim(\mathbf{f}_{v^i}, \mathbf{f}_{v^j})$ is the similarity between two entities' attribute vectors (e.g, it can be calculated by using cosine similarity measure), ϕ is the node type mapping function, $N_{T_{t+1}}(v_t^i)$ denote T_{t+1} type of neighborhood of node v_t^i, v^c denotes a node in $N_{T_{t+1}}(v_t^i)$. Since we have three different meta-paths, we simply combine the path instances sampled via each meta-path, and feed them into the skip-gram model [29] to learn the node embeddings.

2.4 Classification Model

The problem of determining if a given pair of vendors are the same individual can be considered as a link prediction in an AHIN. In this paper, motivated by [1], we propose to devise a classification model for AHIN (named *vIdentifier* as shown in Figure 2.(d)) to predict the likelihood of the link between two nodes.

We first apply a Deep Neural Network (DNN) into node embedding to learn $f_\theta: \mathbb{R}^D \rightarrow \mathbb{R}^d$ that maps a vendor's learned latent representation onto a low-dimensional manifold. The mapping function f_θ is defined as [1]:

$$\begin{aligned} f_\theta: Y_u &\rightarrow FC_{\{\mathbf{W}_1, \mathbf{b}_1\}} \rightarrow BatchNorm \rightarrow relu \\ &\rightarrow FC_{\{\mathbf{W}_2, \mathbf{b}_2\}} \rightarrow BatchNorm \rightarrow f_\theta(Y_u), \end{aligned} \tag{2}$$

where $FC_{\{\mathbf{W}, \mathbf{b}\}}$ is a fully-connected layer with weight matrix $\mathbf{W}$ and bias vector $\mathbf{b}$, BatchNorm is described in [22], $relu(x) = max(0, x)$ is an element-wise activation function, and $\theta = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, ...\}$.

Given a pair of vendors, to define a general link function $g(u, v) \in \mathbb{R}$, we consider a low-rank affine projection in the manifold space [1]: $g(u, v) = f(Y_u)^T \times M \times f(Y_v)$, where the low-rank projection matrix $M = L \times R$ with $L \in \mathbb{R}^{d \times b}$ and $R \in \mathbb{R}^{b \times d}$, $b < d < D$. We can factor $g(u, v)$ into an inner product $\langle L^T f(Y_u), Rf(Y_v) \rangle$.

Then, we utilize the method proposed in [1] to conduct the optimization. The function is devised as following:

$$Pr(\mathcal{G}) \propto \prod_{u,v \in \mathcal{V}} \sigma(g(u, v))^{D_{uv} f(u,v)} (1 - \sigma(g(u, v)))^{1 - f(u,v)}, \tag{3}$$

where $\sigma(x) = 1/(1 + exp(-x))$ is the standard logistic,

$$f(u, v) = \begin{cases} 1 & \text{if } u \text{ and } v \text{ are the same individual} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

D_{uv} is the frequency that vendors u and v co-occur within a specific window in the path instances sampled by our above proposed attribute-aware meta-path random walk.

3 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we conduct three sets of experiments using data collections from darknet markets to fully evaluate the performance of our developed system *uStyle-uID* for drug trafficker identification.

3.1 Experimental Setup

To fully evaluate our proposed method, we have collected the data from four different darknet markets *Valhalla*, *Dream Market*, *SilkRoad2* and *Evolution*. For the former two darknet markets, we develop a set of crawling tools to scrape weekly snapshots from June 2017 to August 2017. For the rest of markets, we collect their public data dumps. After data collection, we merely retain the vendors who at least posted two different drugs, each of which should be at least with one transaction. We summarize the collected data in Table 1. Due to the anonymity of darknet markets, it is difficult to access the actual ground-truth. Similar to the approach used in [40], we propose an alternative solution: for a given vendor, we randomly split his/her posted texts and photos into two even parts to form a positive example; then we randomly match this given vendor to the other in the same darknet market to generate a negative example. In this pseudo setting, the attribute vector attached to each vendor node is set as null. In the following experiments, we conduct ten-fold cross validations and use accuracy (*ACC*) and *F1* as performance measures to evaluate different methods. The parameters for *Vendor2Vec* are empirically set as follows: node dimension $D = 100$, walks per node $r = 10$, walk length $l = 80$ and window size $w = 10$, while for the parameter of d in *vIdentifier* is set as 30 and the remaining ones are consistent with [1].

Table 1: Information of the dataset.

Dataset	Vendors	Drugs	Entities	Relations
Valhalla (Val)	522	13,150	27,535	54,698
Dream Market (DM)	2,547	67,270	139,493	281,080
SilkRoad2 (SR2)	681	15,231	31,354	66,648
Evolution (Evol)	1,650	36,798	79,451	164,792

3.2 Comparisons of Different Features

In this set of experiments, we first evaluate the effectiveness of different features for drug trafficker identification.

- **Text-based features**: (1) *text content only* (***f-1***), (2) *writing style only* (***f-2***), and (3) *text content and writing style* (***f-3***).
- **Photo-based features**: (1) *photo content* (***f-4***), (2) *photography style* (***f-5***), and (3) *photo content and photography style* (***f-6***).

The experimental results are illustrated in Table 2, from which we can see that different features show different performances. To put this into perspective, (1) the relatedness over vendors depicted by style-based correlations (***f-2*** and ***f-5***) perform better than content-based correlations (***f-1*** and ***f-4***); (2) feature engineering (***f-3*** and ***f-6***) helps the performance; (3) the photo-based features (***f-4***, ***f-5***, and ***f-6***) perform better than the text-based ones (***f-1***, ***f-2***, and ***f-3***); (4) our proposed method *uStyle-uID* which integrates different levels of semantics, leveraging both writing and photograph styles, obtains a significantly better performance.

Table 2: Comparisons of different features.

Metric	Method	Feature	Val	DM	SR2	Evol
ACC	Text-based	***f-1***	0.780	0.798	0.782	0.792
		f-2	0.792	0.807	0.795	0.805
		f-3	0.803	0.819	0.806	0.815
	Photo-based	***f-4***	0.795	0.808	0.796	0.805
		f-5	0.807	0.817	0.808	0.816
		f-6	0.818	0.828	0.819	0.827
	uStyle-uID	/	**0.876**	**0.903**	**0.881**	**0.889**
F1	Text-based	***f-1***	0.782	0.784	0.772	0.784
		f-2	0.796	0.788	0.784	0.790
		f-3	0.804	0.809	0.792	0.797
	Photo-based	***f-4***	0.785	0.794	0.792	0.793
		f-5	0.798	0.806	0.796	0.802
		f-6	0.810	0.817	0.806	0.809
	uStyle-uID	/	**0.865**	**0.894**	**0.868**	**0.879**

3.3 Network Embedding Model Comparisons

In this set of experiments, we evaluate our proposed method *Vendor2Vec* by comparisons with several state-of-the-art network embedding models including **DeepWalk** [31], **node2vec** [18] and **metapath2vec** [10]. For these embedding methods, we use the same parameters as *Vendor2Vec*. The results from Table 3 show that *Vendor2Vec* consistently and significantly outperforms all state-of-the-art embedding models. The success of *Vendor2Vec* lies in: (1) the proper consideration and accommodation of the heterogeneous property of AHIN; (2) the advantage of the attribute setting and the proposed attribute-aware meta-path guided random walk for sampling the high-quality path instances (i.e., without the attribute information such as writing and photography styles, the generated path instances are of low quality and less useful to our application).

Table 3: Comparisons of network embedding models.

Metric	Method	Val	DM	SR2	Evol
ACC	DeepWalk	0.703	0.714	0.699	0.707
	node2vec	0.726	0.731	0.722	0.729
	metapath2vec	0.741	0.754	0.744	0.748
	Vendor2Vec	**0.876**	**0.903**	**0.881**	**0.889**
F1	DeepWalk	0.682	0.688	0.669	0.679
	node2vec	0.703	0.710	0.692	0.709
	metapath2vec	0.718	0.735	0.727	0.731
	Vendor2Vec	**0.865**	**0.894**	**0.868**	**0.879**

3.4 Comparisons with Alternative Approaches

In this set of experiments, based on our collected datasets, we compare our developed system *uStyle-uID* with alternative approaches:

(1) feeding all the features (i.e., ***f-3***, ***f-6***, and feature vectors of vendors and drugs) into a generic DNN [19] to make the identification (denoted as *Hybrid-DNN*); (2) replacing the *vIdentifier* in *uStyle-uID* by a generic DNN (denoted as *AHIN-DNN*); (3) replacing the *vIdentifier* in *uStyle-uID* by SVM (denote as *AHIN-SVM*). For the generic DNN, we implement the model in Keras [40] and retain the default parameters. The experimental results are illustrated in Table 4. From the results we can observe that *AHIN-DNN* added the knowledge represented as AHIN performs better than *Hybrid-DNN*, which shows that using meta-path based approach over AHIN is able to build the higher-level semantic connection between vendors with a more expressive view. We also note that *uStyle-uID* significantly outperforms other baselines, which demonstrates that *uIdentifier* indeed helps the performance compared with the generic DNN and state-of-the-art shallow learning classification model.

Table 4: Comparisons of other alternative approaches.

Metric	Method	Val	DM	SR2	Evol
ACC	Hybrid-DNN	0.831	0.839	0.833	0.838
	AHIN-DNN	0.854	0.876	0.856	0.863
	AHIN-SVM	0.843	0.851	0.847	0.848
	uStyle-uID	**0.876**	**0.903**	**0.881**	**0.889**
F1	Hybrid-DNN	0.809	0.818	0.812	0.814
	AHIN-DNN	0.841	0.864	0.845	0.855
	AHIN-SVM	0.832	0.837	0.832	0.840
	uStyle-uID	**0.865**	**0.894**	**0.868**	**0.879**

4 CROSS-MARKET DRUG TRAFFICKER IDENTIFICATION AND CASE STUDIES

To better understand and gain deeper insights into the ecosystem of drug trafficking in darknet markets, we further apply our developed system *uStyle-uID* for cross-market drug trafficker identification. For the detected cross-market vendor pairs, we further sample 798 pairs and validate them using conclusive evidences. Among these 798 detected cross-market pairs, 726 pairs (90.09%) are with high confidence that they are the same individuals and 22 pairs are uncertain (2.76%). As shown in Figure 4, for one of our detected cross-market vendor pairs, though "The***shop" on Evolution and "***Store" on Valhalla have different usernames, PGP keys and contact information, they sell similar drugs (i.e., heroin and MDMA) with similar writing and photography styles. After further investigation , we find that "The***shop" and "***Store" are both the members of a group named "***DrugShop" in Finland according to the description in the terms and conditions of vendors. This indicates that they might work together as an organization and we can link them to the same group of drug traffickers. Such kind of information can provide investigative insight for law enforcement to trace their activities and thus to build proactive defenses.

5 RELATED WORK

To combat drug trafficking in darknet markets, there have been many research efforts on darknet market data analysis [3, 9, 20, 40]. Among these studies, there have been some methods proposed

(a) Vendor "The***shop" on Evolution (b) Vendor "***Store" on Valhalla

Figure 4: An example of detected cross-market vendor pair.

to tackle the challenges of authorship identification such as [20, 40]. However, these approaches mainly relied on either stylometry analysis or photography style analysis. Different from existing works, we propose to leverage both writing and photography styles together with their contents for drug trafficker identification.

In order to depict different entities, associated attributes and the rich relationships among them, it is important to model them properly. Though HIN has shown the success of modeling different types of entities and relations [11–13, 21, 35, 42, 43], it has limited capability of modeling additional attributes attached to entities. To address this challenge, we propose to use AHIN for representation. To better address representation learning for HIN, many efficient network embedding methods have been proposed such as metagraph2vec [11], metapath2vec [10], HIN2vec [14]. However, these models are unable to deal with the attribute information associated with each entity. To address this issue, we propose *Vendor2Vec* to learn the desirable node representations in AHIN.

6 CONCLUSION

To combat drug trafficking, in this paper, we design and develop an intelligent system named *uStyle-uID* to automate drug trafficker identification in darknet markets. In *uStyle-uID*, we propose to leverage both writing and photography styles at the first attempt. To depict vendors, drugs, texts, photos and their associated attributes as well as the rich relationships among them, we present a structural AHIN to model them which gives the vendors higher-level semantic representations. Then, a meta-path based approach is used to characterize the semantic relatedness over vendors. To efficiently measure the relatedness over vendors in AHIN, we propose a new network embedding model *Vendor2Vec* which leverages complementary attribute information attached in the nodes to guide the meta-path based random walk for path instances sampling. Then, we transform the identification task into a link prediction problem and further present a learning model named *vIdentifier* to solve the problem. The promising experimental results on the collected datasets from four darknet markets demonstrate that *uStyle-uID* outperforms alternative approaches.

7 ACKNOWLEDGMENTS

Y. Zhang, Y. Fan, W. Song, S. Hou, Y. Ye and X. Li's work is partially supported by the DoJ/NIJ under grant NIJ 2018-75-CX-0032; the NSF under grants CNS-1618629, CNS-1814825, CNS-1845138 and OAC-1839909; the WV Higher Education Policy Commission Grant (HEPC.dsr.18.5); and the WVU Research and Scholarship Advancement Grant (R-844).

REFERENCES

[1] Sami Abu-El-Haija, Bryan Perozzi, and Rami Al-Rfou. 2017. Learning edge representations via low-rank asymmetric projections. In *CIKM*. ACM, 1787–1796.

[2] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger finder: Taking stylometry to the underground. In *Security and Privacy (SP), 2014 IEEE Symposium on*. IEEE, 212–226.

[3] Julian Broséus, Damien Rhumorbarbe, Caroline Mireault, Vincent Ouellette, Frank Crispino, and David Décary-Hétu. 2016. Studying illicit drug trafficking on Darknet markets: structure and organisation from a Canadian perspective. *Forensic science international* 264 (2016), 7–14.

[4] Julia Buxton and Tim Bingham. 2015. The rise and challenge of dark net drug markets. *Policy Brief* 7 (2015).

[5] Prima Chairunnanda, Nam Pham, and Urs Hengartner. 2011. Privacy: Gone with the Typing! Identifying Web Users by Their Typing Patterns.. In *SocialCom/PASSAT*. 974–980.

[6] Richard Colbaugh and Kristin Glass. 2011. Proactive defense for evolving cyber threats. In *ISI*. IEEE, 125–130.

[7] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*. Springer, 288–301.

[8] Hakan Demirbuken and others. 2011. The Global Afghan Opium Trade: A Threat Assessment. *UNODC Report* (2011), 3–4.

[9] Martin Dittus, Joss Wright, and Mark Graham. 2018. Platform Criminalism: The'last-mile'geography of the darknet market supply chain. In *WWW*. International World Wide Web Conferences Steering Committee, 277–286.

[10] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*. ACM, 135–144.

[11] Yujie Fan, Shifu Hou, Yiming Zhang, Yanfang Ye, and Melih Abdulhayoglu. 2018. Gotcha-sly malware!: Scorpion a metagraph2vec based malware detection system. In *KDD*. ACM, 253–262.

[12] Yujie Fan, Yiming Zhang, Yanfang Ye, and Xin Li. 2018. Automatic Opioid User Detection from Twitter: Transductive Ensemble Built on Different Meta-graph Based Similarities over Heterogeneous Information Network.. In *IJCAI*. 3357–3363.

[13] Yujie Fan, Yiming Zhang, Yanfang Ye, Wanhong Zheng, and others. 2017. Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from twitter and case studies. In *CIKM*. ACM, 1259–1267.

[14] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In *CIKM*. ACM, 1797–1806.

[15] Dario Garcia-Gasulla, Eduard Ayguadé, Jesús Labarta, Javier Béjar, Ulises Cortés, Toyotaro Suzumura, and R Chen. 2017. A visual embedding for the unsupervised extraction of abstract semantics. *Cognitive Systems Research* 42 (2017), 73–81.

[16] Michael Gilbert and Nabarun Dasgupta. 2017. Silicon to syringe: Cryptomarkets and disruptive innovation in opioid supply chains. *International Journal of Drug Policy* 46 (2017), 160–167.

[17] P. Griffiths and J. Mounteney. 2017. Disruptive potential of the internet to transform illicit drug markets and impact on future patterns of drug consumption. *Clinical Pharmacology & Therapeutics* 101, 2 (2017).

[18] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*. ACM, 855–864.

[19] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and others. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.

[20] Thanh Nghia Ho and Wee Keong Ng. 2016. Application of stylometry to darkweb forum user identification. In *ICS*. Springer, 173–183.

[21] Shifu Hou, Yanfang Ye, Yangqiu Song, and Melih Abdulhayoglu. 2017. Hindroid: An intelligent android malware detection system based on structured heterogeneous information network. In *KDD*. ACM, 1507–1515.

[22] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[23] Dev Kar and Joseph Spanjers. 2017. Transnational crime and the developing world. (2017).

[24] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *CVPR*, Vol. 1. IEEE, 419–426.

[25] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*. 1188–1196.

[26] Xiang Li, Yao Wu, Martin Ester, Ben Kao, Xin Wang, and Yudian Zheng. 2017. Semi-supervised clustering in attributed heterogeneous information networks. In *WWW*. 1621–1629.

[27] Dream Market. 2018. *Dream Market*. http://6khhxwj7viwe5xjm.onion.

[28] James Martin. 2014. Drugs on the dark net: How cryptomarkets are transforming the global trade in illicit drugs. (2014).

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[31] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *KDD*. ACM, 701–710.

[32] Damien Rhumorbarbe, Ludovic Staehli, Julian Broseus, Quentin Rossy, and Pierre Esseiva. 2016. Buying drugs on a Darknet market: A better deal? Studying the online illicit drug market through the analysis of digital, physical and chemical data. *Forensic science international* 267 (2016), 173–182.

[33] SilkRoad. 2018. *SilkRoad*. http://silkroad7rn2puhj.onion.

[34] Martijn Spitters, Femke Klaver, Gijs Koot, and Mark van Staalduinen. 2015. Authorship analysis on dark marketplace forums. In *EISIC*.

[35] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB Endowment* 4, 11 (2011), 992–1003.

[36] Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32, 5 (1998), 323–352.

[37] Valhalla. 2018. *Valhalla*. http://valhallaxmn3fydu.onion.

[38] Joe Van Buskirk, Amanda Roxburgh, Michael Farrell, and Lucy Burns. 2014. The closure of the Silk Road: What has this meant for online drug trading? *Addiction* 109, 4 (2014).

[39] Joe Van Buskirk, Amanda Roxburgh, Sundresan Naicker, and Lucinda Burns. 2015. A response to Dolliver's "Evaluating drug trafficking on the Tor network". *International Journal of Drug Policy* 26, 11 (2015).

[40] Xiangwen Wang, Peng Peng, Chun Wang, and Gang Wang. 2018. You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces. In *ACCCS*. ACM, 431–442.

[41] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*. 649–657.

[42] Yiming Zhang, Yujie Fan, Shifu Hou, Jian Liu, Yanfang Ye, and Thirimachos Bourlai. 2018. iDetector: Automate Underground Forum Analysis Based on Heterogeneous Information Network. In *ASONAM*. IEEE, 1071–1078.

[43] Yiming Zhang, Yujie Fan, Yanfang Ye, Liang Zhao, Jiabin Wang, Qi Xiong, and Fudong Shao. 2018. KADetector: Automatic Identification of Key Actors in Online Hack Forums Based on Structured Heterogeneous Information Network. In *ICBK*. IEEE, 154–161.